INDEX

Abstract

Area of topic : Oncology , Prediction based modelling using graph mining

ABSTRACT :

Among thousands of potential mutations, identifying and separating cancer driver genes remains a big difficulty. Precise identification of driver genes and mutations is crucial for cancer research and treatment personalization based on accurate patient classification. Many driver mutations within a gene exist at low rates due to inter-tumor genetic variability, making it difficult to identify them from non-driver mutations.

This project uses a transcription adjustment network and its data set from the database REGNETWORK. The subject of our project is the discovery of genes that cause lung cancer with a network approach. To do this, centralization and socialization in graph are used. The degree of centrality, degree of mediocrity, and proximity are considered as parameters in identifying lung cancer gene (with the cancer-causing mutation). Socializing of data is done to find genes that are more closely related to each other. Transcriptional regulatory network is a type of biological network that is made up of different transcription factors and genes and their interactions. We Analyse these networks to examine the flow of information in a biological system and identify pathways that are useful for different functions. Nodes in this network are genes and transcripts, so there are two types of modules in the network, Gene module and transcription factor module. Edges represents physical or regulatory interaction between them. We use 2 different algorithms to build the network model and comparison is done using Accuracy,F1 Score, Recall, Precision.

KEYWORDS : Graph mining, TRNs, driver-gene, Integration-matrix, Network socialization

CHAPTER 1

DATA SET:

Database URL: http://www.regnetworkweb.org

Transcriptional and post-transcriptional regulation of gene expression is of fundamental importance to numerous biological processes. Nowadays, an increasing number of gene regulatory relationships have been documented in various databases and literature. However, to more efficiently exploit such knowledge for biomedical research and applications, it is necessary to construct a genome-wide regulatory network database to integrate the information on gene regulatory relationships that are widely scattered in many different places. Therefore, in this work, we used a knowledge-based database, named 'RegNetwork', of gene regulatory networks for human and mouse by collecting and integrating the documented regulatory interactions among transcription factors (TFs), microRNAs (miRNAs) and target genes from 25 selected databases.

RegNetwork is a data repository of five-type transcriptional and posttranscriptional regulatory relationships for human and mouse:

TF→TF

TF→gene

TF→miRNA

miRNA→TF

miRNA→gene

RegNetwork integrates the curated regulations in various databases and the potential regulations inferred based on the transcription factor binding sites (TFBSs). Transcription factor (TF) and microRNA (miRNA) are central regulators in gene regulations. They function in the transcriptional and posttranscriptional levels respectively. Recently, more and more regulatory relationships in databases and literatures are available. It will greatly valuable for studying gene regulatory systems by integrating the prior knowledge of the transcriptional regulations between TF and target genes, and the posttranscriptional regulations between miRNA and targets. The conservation knowledge of transcription factor binding site (TFBS) can also be implemented to couple the potential regulatory relationships between regulators and their targets.

From RegNetwork, we can query and identify the combinatorial and synergic regulatory relationships among TFs, miRNAs and genes.

It is a database of transcriptional and posttranscriptional regulatory networks in human and mouse. TF and miRNA are two major regulators controlling gene expression. RegNetwork collects the knowledge-based regulatory relationships, as well as some potentially regulatory relationships between the two regulators and targets. It provides a platform of depositing the known and predicted gene regulations in the transcriptional and posttranscriptional levels simultaneously. The knowledge-derived regulatory networks are expected to be greatly beneficial for identifying critical regulatory programs in various context-specific conditions.

This project uses a transcription adjustment network and its data set from the database.

Element	Description	Number	
		Human	Mouse
Node	All nodes included in the regulatory network	23 079	20 738
Edge	All regulatory relationships included in the regulatory network	369 277	323 636
TF	The documented TFs included in the regulatory network	1456	1328
miRNA	The miRNAs included in the regulatory network	1904	1290
Gene	The target genes included in the regulatory network	19 719	18 120
TF–gene	The 'TF–gene' regulations included in the regulatory network	149 841	94 876
TF–TF	The 'TF'–'TF gene' self-regulations included the regulatory network	361	129
TF–miRNA	The 'TF–miRNA gene' regulations included in the regulatory network	21 744	25 574
miRNA–gene	The 'miRNA–target gene' regulations included in the regulatory network	171 477	176 512
miRNA–TF	The 'miRNA–TF gene' regulations included in the regulatory network	25 854	26 545

Fig 1 The basic statistics of the regulatory networks of human and mouse in RegNetwork.

We recovered the information related to the human TRN. Of course, in this network, there was some other information about interactions in the regulatory network including the regulatory network of miRNA on that were removed from the final network being studied. The final information included 150202 interactions related to TF-TF and TF-mRNA. In this network, for each interaction, a confidence level has reported. We used it to assign a weight to each interaction. These confidence levels included “low”, “high” and medium. Assigned edge weights According to previous research (0.2, 0.5 and 0.8 for low, medium and high confidence).

The first 10 data are shown in Table 1, which is a column. The target node is Target and the source node represents the weight of the network edges p The relationships between these nodes are considered directional, so the resulting graph is directional graph, and by performing the final prerecessions, the graph with 87388 edges and 11016 knots and the average value of internal and external degrees is 7.9328 and this information is shown in Table 2.

	Target	Source	p
0	ABL1	SHC3	0.8
1	ABL1	STAT5B	0.8
2	ABL1	CBLB	0.8
3	ABL1	CBLC	0.8
4	ABL1	CD55	0.8
5	ABL1	CRK	0.8
6	ABL1	CRKL	0.8
7	ABL1	RAC3	0.8
8	ABL1	RB1	0.8
9	ABL1	SHC1	0.8

Table 1: Overview of the final data used

```
Name:
Type: DiGraph
Number of nodes: 11016
Number of edges: 87388
Average in degree:   7.9328
Average out degree:   7.9328
```

Table 2: Information of the desired network

NETWORK ILLUSTRATION

The general shape of the transcription regulation network and its modules is as shown in Figures 3 and 4, here are the relevant data. Released miRNA

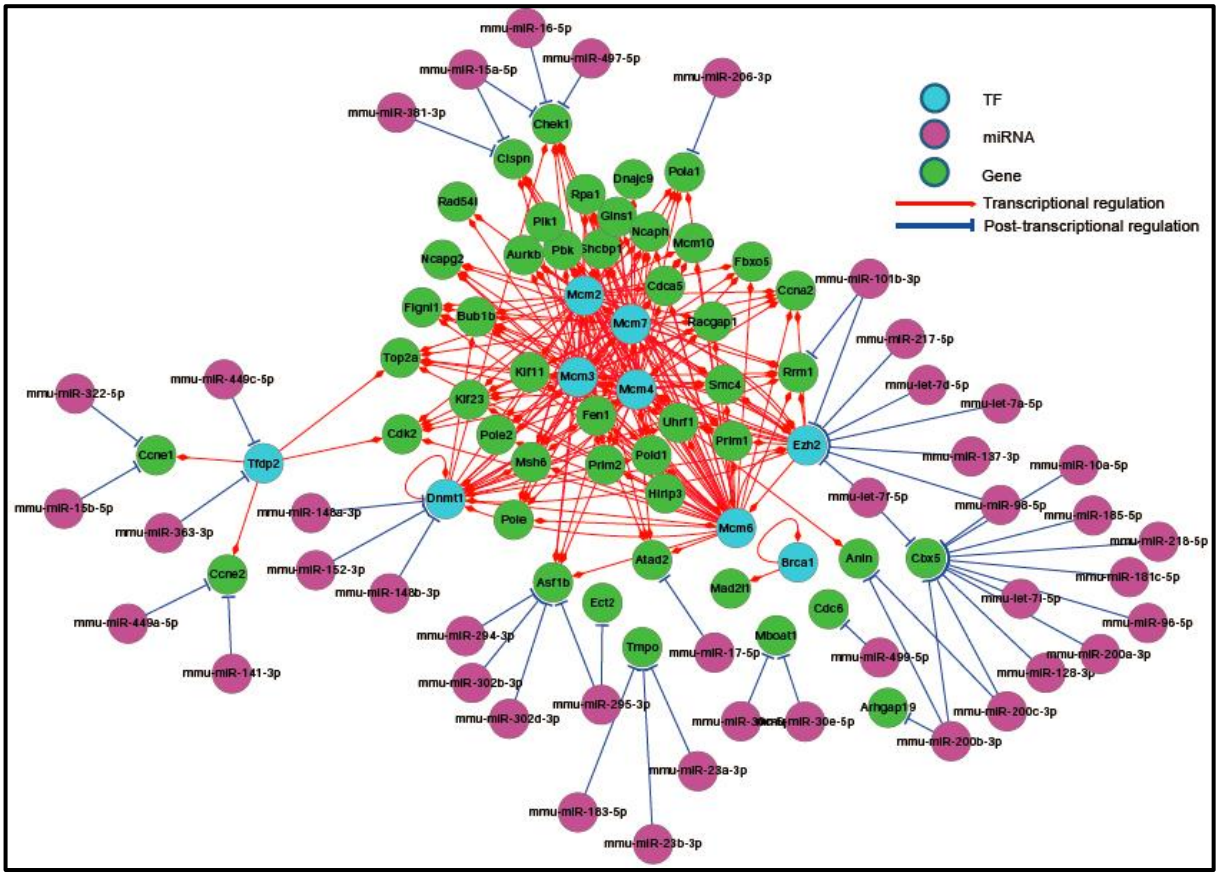


Figure 4: Transcription adjustment grid on a smaller scale

Six siRNAs targeted the six important immunological genes Ifnb1, Irf3, Irf5, Stat1, Stat2 and Nfkb2 transcripts respectively in primary mouse macrophages. They profiled the genome-wide gene expressions and identified several co-expressed gene clusters. By RegNetwork, we can easily extract the existing regulatory relationships in each gene cluster respectively. Figure 2 illustrates some of the regulatory relationships in such a cluster (The third gene cluster in the original paper). Moreover, RegNetwork identifies the miRNA regulatory relationships with these genes simultaneously.

Compared to the original gene set, the regulatory wiring information directs further interesting analyses and experimental designs about the influences transmitted between these genes in response to the siRNA treatment. RegNetwork provides a resource for depositing the existing TF and miRNA mediated regulations, which are expected to benefit many regulatory researches and experimental designs.

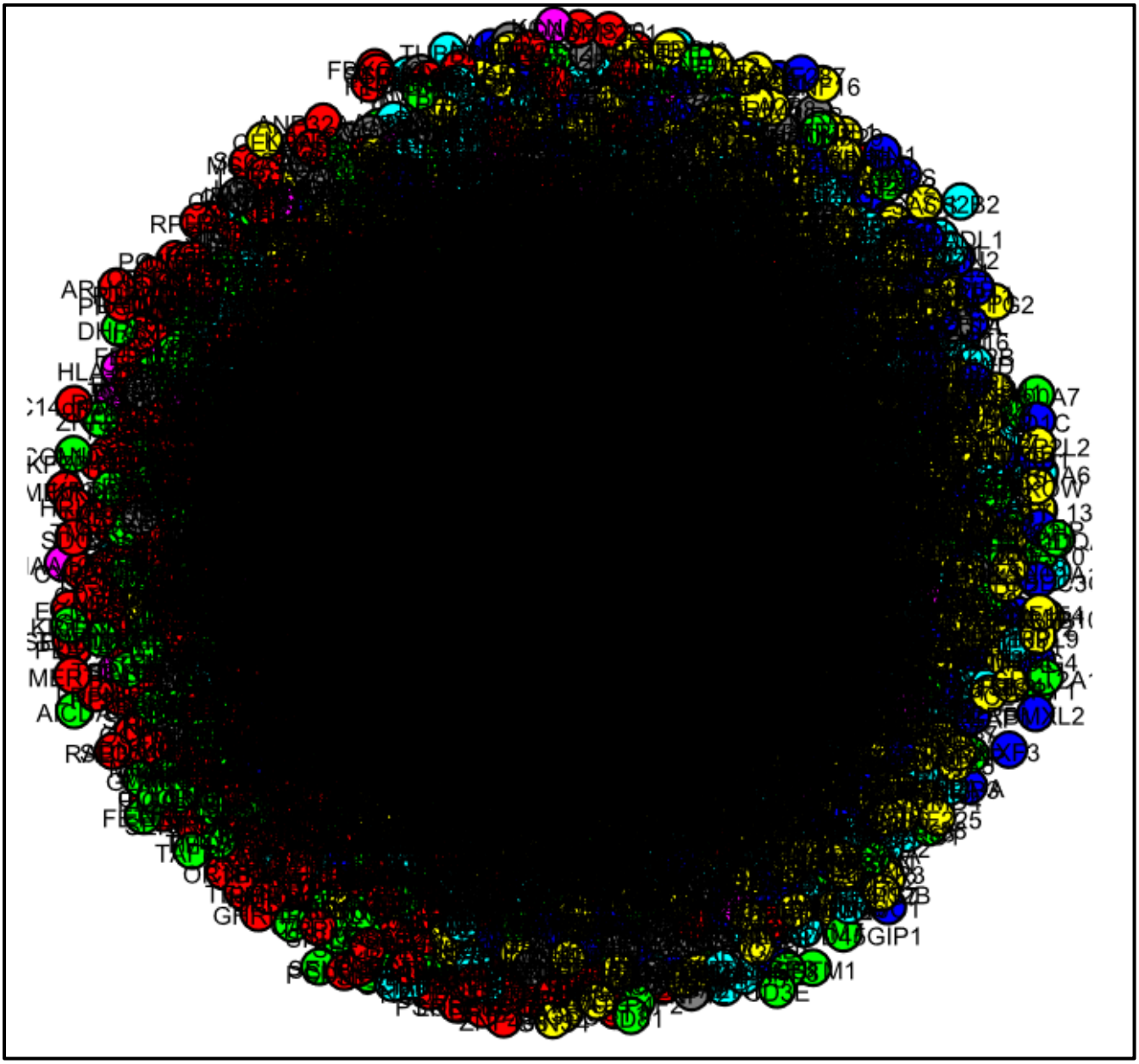


Figure 2: Illustration of the desired network using an algorithm force - directed

The considered graph is weighted, oriented and, as shown in Figure 2, uncoordinated, and for network analysis because in most Sometimes the condition of being connected is necessary (such as calculating the proximity centres, intermediate, etc.) first turn it into a link and connect it on the network Analyses are performed. The network consists of 2 weak link components and 9997 strong link components. Use the maximum weak link component to network the information in this graph can be seen in Table 6.

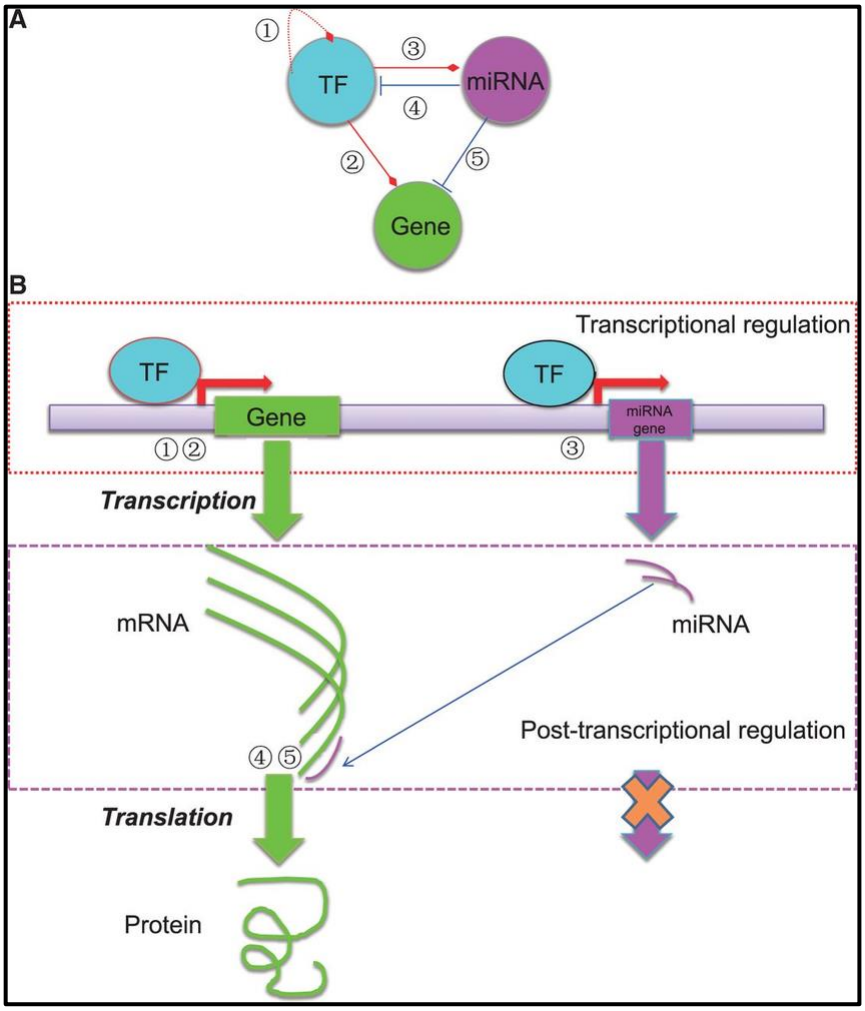


Figure 3: Large-scale transcription adjustment network

Depending on the number of nodes (only one node, node Deleted) and the type of nodes (node mentioned as non-cancerous and insignificant in this matter, ZNF471 The resulting network is suitable for analysis.

Literature Review:

Objective	Data	Method	Description
Recurrent somatic mutation identification	SNV	MutSigCV [48]	Uses coverage information and genomic features (e.g. DNA replication time) to estimate the background mutation rate of a gene.
		MuSiC [49]	Uses a per-gene background mutation rate; allows for user-defined regions of interest.
		Youn *et al.*[51]	Includes predicted impact on protein function in determining recurrent mutations.
		Sjöblom *et al.*[52]	Defines a cancer mutation prevalence score for each gene.
		DrGaP [139]	Uses Bayesian approach to estimate background mutation rate; helpful for cancer types with low mutation rate.
	CNA	GISTIC2 [61], JISTIC [63]	Uses 'peel-off' techniques to find smaller recurrent aberrations inside larger aberrations.
		CMDS [62]	Identifies recurrent CNAs from unsegmented data.
		ADMIRE [65]	Multi-scale smoothing of copy number profiles.

Functional impact prediction	General	SIFT [72]	Uses conservation of amino acids to predict functional impact of a non-synonymous amino-acid change.
		Polyphen-2 [74]	Infers functional impact of non-synonymous amino-acid changes through alignments of related peptide sequences and a machine-learning-based probabilistic classifier.
		MutationAssessor [75]	Uses protein homologs to calculate a score based on the divergence in conservation caused by an amino-acid change.
		PROVEAN [73]	Benchmarks favorably against MutationAssessor, Polyphen-2 and SIFT.
	Cancer-specific	CHASM [77]	Uses a machine-learning approach to classify mutations as drivers or passengers based on sequence conservation, protein domains, and protein structure.
		Oncodrive-FM [79]	Combines scores from SIFT, Polyphen-2, and MutationAccessor into a single ranking.
	Positional or structural clustering	NMC [83]	Finds clusters of non-synonymous mutations across patients. Typically used with missense mutations to detect so-called 'activating' mutations.

		iPAC [84]	Extends the NMC approach to search for clusters of mutations in three-dimensional space using crystal structures of proteins.
Pathway analysis and combinations of mutations	Known pathways	GSEA [92]	A general technique for testing ranked lists of genes for enrichment in known gene sets. Can be used on rankings derived from significance of observed mutations.
		PathScan [95]	Finds pathways with excess of mutations in a gene set (pathway), by combining *P*-values of enrichment across samples.
		Patient-oriented gene sets [94]	Tests known pathways using a binary indicator for a pathway in each patient.
	Interaction networks	NetBox [140]	Finds network modules in a user-provided list of genes. Significance depends only on the topology of the genes in the network, and not on mutation scores.
		HotNet [102]	Finds subnetworks with significantly more aberrations than would be expected by chance, using both network topology and user-defined gene or protein scores.
		MEMo [104]	Finds subnetworks whose interacting pairs of genes have mutually exclusive aberrations

			[105]; recommends including only recurrent SNVs and CNAs in the analysis.
	De novo	Dendrix [102]	Identifies groups of genes with mutually exclusive aberrations.
		Multi-Dendrix [112]	Simultaneously finds multiple groups of genes with mutually exclusive aberrations.
		RME [110]	Finds groups of genes with mutually exclusive aberrations by building from gene pairs; best results obtained when restricting to genes with high mutation frequencies (e.g. > 10%).

Chapter 2

Transcriptional initiation and elongation

Transcriptional initiation and elongation, RNA stability, and accessibility and rate of translation are all examples of processes where gene expression can be controlled. The action of transcription factors (TFs) and cis-regulatory DNA regions regulates transcriptional start in our experiment. TRNs are TF-target gene regulatory interactions. TRN edges denote direct connections between a transcription factor and its target genes. TRN models are systems-level models that describe developmental and physiological activities. TRN information is useful in a variety of basic and applied biomedical researches. It has the potential to improve our understanding of the molecular principles of development and cellular reprogramming, resulting in more efficient methods for creating various cell types for regenerative therapies. TRN dysfunction-related illnesses have mechanisms that can be explored. TRN information can help with the development of effective cellular engineering technologies and the selection of novel pharmaceutical targets.

TRNs are extremely complicated, as seen by the large number of regulatory components and intricate connectivity patterns amongst them. In reaction to external or internal cues, they usually display highly dynamic and often nonlinear behaviours. As a result, computer modelling is an important part of TRN research.

High-throughput technologies have substantially increased our ability to collect complementing data sources for TRN computational modelling in recent years. Using genome-scale data sets, we investigated various computational approaches for inferring TRN models. TRNs inferred from large-scale data are less precise, but they can be used to infer network components and wiring, which is especially important for TRNs that are mostly uncharacterized. Then, in both normal and pathological development, we outline representative TRNs created utilising large-scale techniques. Then, based on a review of numerous large-scale TRN models, we provide new insights into the structure/function relationship of TRNs. Finally, we discuss some unanswered problems about TRNs, such as how to integrate heterogeneous data types to increase model accuracy, how to infer condition-specific TRNs, and how to compare TRNs across species and situations to better understand their structure/function relationship.

Computational approaches for constructing genome-wide TRN models

As illustrated in figure 1, current techniques to building computational models of TRNs can be divided into three categories based on the type of data used for inference. The only input for the first class is gene expression data. This group of methods is known as the reverse engineering approach since they start with the regulatory output (e.g., expression level). A variety of computational frameworks have been used to build approaches in this class, including linear regression, statistical correlation, and Bayesian networks. The underlying premise behind regression-based techniques is that the expression levels of TFs that directly regulate a target gene are the most informative among all TFs for predicting the target gene's expression level. Non-zero regression coefficients show statistical dependency, which is regarded as a regulatory interaction between the TF and the gene when the expression level of a target gene is regressed on the expression levels of TFs. Because there are so many potential TFs to include in a regression, a feature selection approach using regularised regression techniques is usually used to find the regulatory TFs.

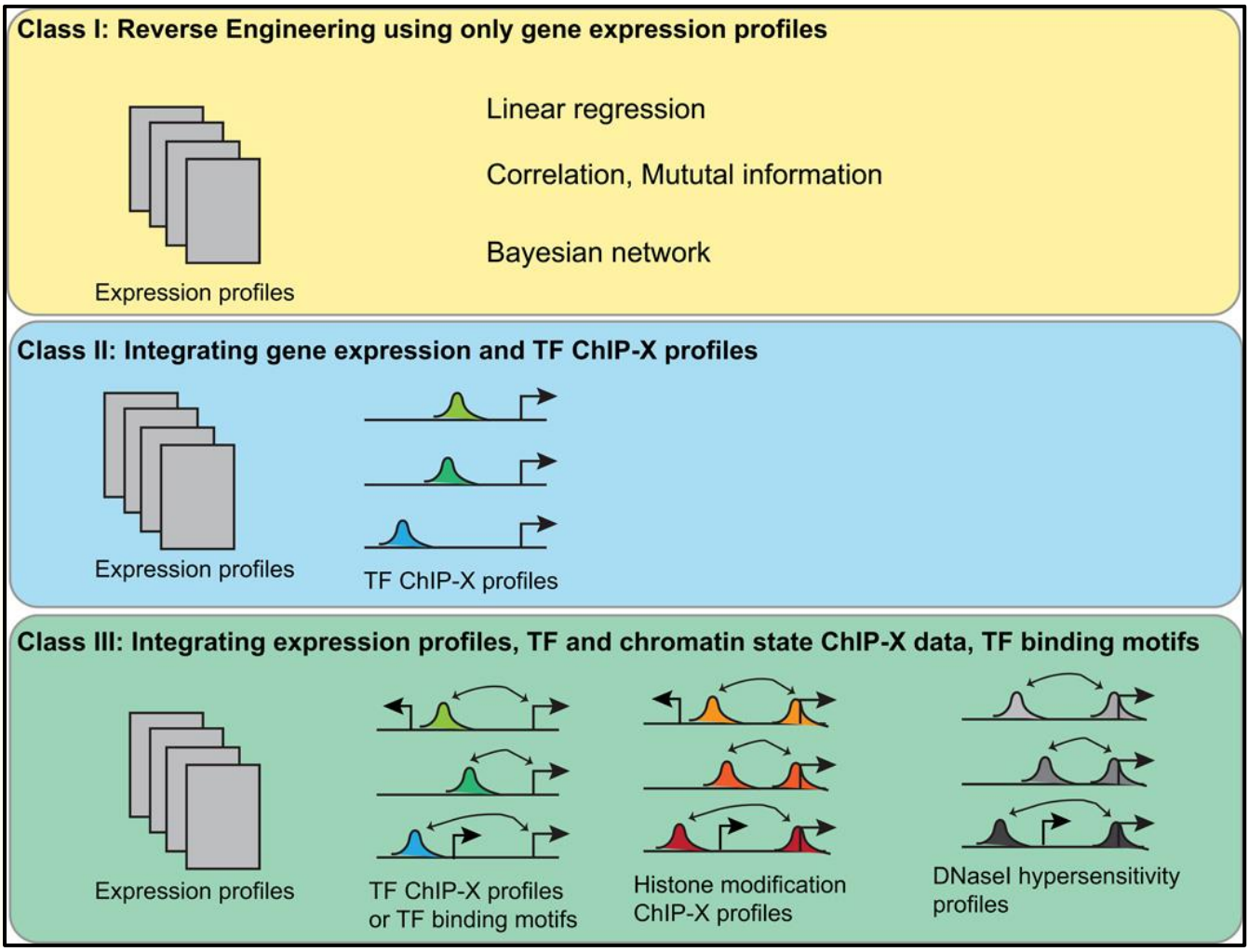


Figure4: Inferring transcriptional regulatory networks can be done using three types of computational approaches.

The protocol ChIP-X stands for ChIP-ChIP or ChIP-Seq. arcs across genomic loci and transcription start sites indicate enhancer-promoter linkages in the panel exhibiting class III approaches.

TRNs in diseased cells:

Given the importance of transcriptional regulation in development and cellular homeostasis, it's no surprise that TRN disruption can result in a variety of illnesses. Mutations in regulatory DNA sequences, transcription factors, co-factors, and chromatin regulators are examples of such disturbances. The underlying disease-specific TRNs in which the mutant factors work must be characterised in order to understand the roles of mutations in pathogenesis. TRNs underlying numerous human disorders, including cancer, are increasingly being studied using the same high throughput technology used to investigate TRNs in normal cells. To promote pathogenesis, the same

TRNs in normal cells are either rewired or gain changed activity in sick cells, according to new research.

TRNs have been created in cancer cells using reverse engineering techniques. Carro et al. created a glioma-specific TRNs and discovered two TFs (C/EBP and STAT3) that work together to promote neoplastic transformation. In T cell acute lymphoblastic leukaemia, Gatta et al. unravel the carcinogenic TRNs controlled by the two TFs TLX1 and TLX3 (T-ALL).

As ChIP-X technology grew in sophistication and sensitivity, it was increasingly employed to map TRNs in cancer cells. In T-ALL, the oncogenic TF TAL1 creates an integrated autoregulatory loop with two TF partners (RUNX1 and GATA3), according to a recent study. This circuitry is involved in the long-term activation of the TAL1-controlled oncogenic programme. Chromosome translocation events produce a large number of oncogenic TFs. The RUNX1/ETO fusion TF, which is produced by the chromosomal translocation t, is a well-known example. Ptasinska et al. demonstrated that the transcriptional programme driving leukemic proliferation is governed by a dynamic equilibrium between TRNs regulated by the RUNX1/ETO fusion TF and intact RUNX1 complexes using ChIP-Seq and expression profiling. TRNs in breast cancer, prostate cancer, and lung cancer have all been examined using the same method.

Structure/function relationship of TRNs:

Previous efforts to map TRNs (on a local and big scale) have resulted in a vast number of TRN models. The following organisational concepts emerged from an examination of available TRN architectures. TRNs have a worldwide hierarchical topology, for starters. TRNs for embryonic development of animal body parts, such as those for endomesoderm specification in sea urchins, gut and mesoderm determination in C. elegans, and eye lens field specification in zebra fish, exhibit a deep hierarchical architecture. TRNs for terminal fate selection from multipotent stem cells and physiological responses, such as the bifurcation of erythroid versus myeloid fates, the diversification of T helper versus killer cells, and the innate immune response, are rather shallow in comparison. The difference in structural requirements between embryonic development TRNs and terminal fate TRNs reflects the functional

requirements. The development of the body plan necessitates a long series of progressive decisions in several spatiotemporal domains. Terminal cell destiny, on the other hand, can be determined with fewer choices.

Many over-represented tiny connection patterns, known as network motifs, are embedded in the global hierarchical topology of TRNs. Different types of network motifs are distinguished by their distinct connection patterns and associated functionalities. Positive and negative feedback loops are some of the most basic and common patterns. Systems with switch-like behaviour, memory, or biostability frequently have positive feedback loops. Systems with considerable noise resistance to disturbances are functionally related with negative feedback loops. Feedforward loop is a significantly more complicated theme (FFL). This pattern is made up of three genes: a transcription factor (TF), gene X (which regulates TF Y), and gene Z (which is regulated by both X and Y). The net sign of the regulatory actions of the two arms of the motif divides FFL motifs into two types. In coherent FFL motifs, both arms have the same net sign of actions, whereas incoherent FFL motifs have different net signs of actions. It has been demonstrated that coherent FFLs can filter out short spurious signal pulses. Gene Z, instead of responding to a transient signal, only responds to a persistent signal that is above its threshold. The two arms of the FFL act in conflict in incoherent FFLs. After X is triggered, Z has been demonstrated to induce pulse-like response dynamics.

Transcription:

The process of copying (transcription) a gene's DNA sequence into an RNA molecule is known as transcription. Transcription is an important step in converting genetic information into protein. If you're new to those concepts, Sal's central doctrine video is a good place to start. When a gene in DNA is "switched on," or used to generate the protein it specifies, it is called gene expression. Not all of our genes are activated at the same time, in the same cells, or in the same cells or parts of the body.

• If a gene isn't transcribed in a cell, it can't be used to build a protein in that cell.

• If a gene gets transcribed, it will almost certainly be employed to build a protein (expressed). In general, the more a gene is transcribed, the more protein that will be made.

A number of factors influence how much a gene is transcribed. For example, the availability of a gene for transcription can be influenced by how tightly the gene's DNA is twisted around its supporting proteins to form chromatin. Transcription factors, on the other hand, play a particularly important role in transcription regulation. These crucial proteins aid in determining which genes are active in each of your body's cells.

Transcription factors:

The RNA polymerase enzyme, which creates a new RNA molecule from a DNA template, must bind to the gene's DNA. The promoter is where it connects to the body.

RNA polymerase connects directly to the promoter DNA in bacteria. The lac operon and trp operon films show how this mechanism works and how transcription factors can regulate it.

There is an additional stage in humans and other eukaryotes. Only proteins known as basal (general) transcription factors allow RNA polymerase to connect to the promoter. They are essential for the transcription of any gene and are part of the cell's core transcription toolkit.

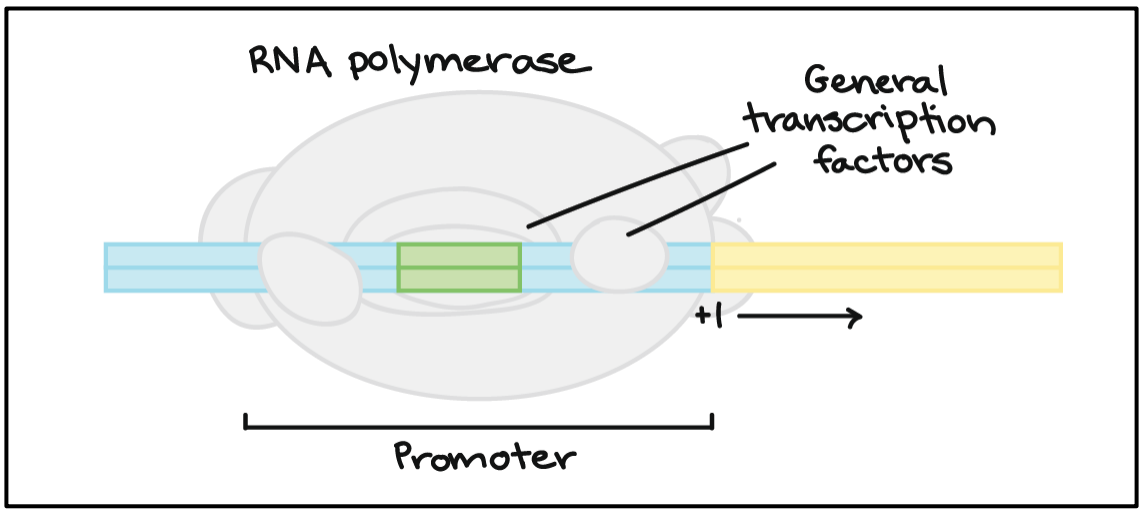


Fig: With the help of a group of proteins known as general transcription factors, RNA polymerase binds to a promoter.

Many transcription factors, including some of the most interesting, are not of the general type. Instead, there is a broad family of transcription factors that regulate the expression of certain genes. For example, a transcription factor may only activate a subset of genes required in specific neurons.

WORKING OF TRANSCRIPTION FACTORS

A typical transcription factor binds to a specific sequence of DNA. Once bound, the transcription factor makes RNA polymerase's binding to the gene promoter either harder or easier.

Activators:

Transcription is triggered by some transcription factors. As indicated in the picture below, they may aid in the binding of general transcription factors and/or RNA polymerase to the promoter.

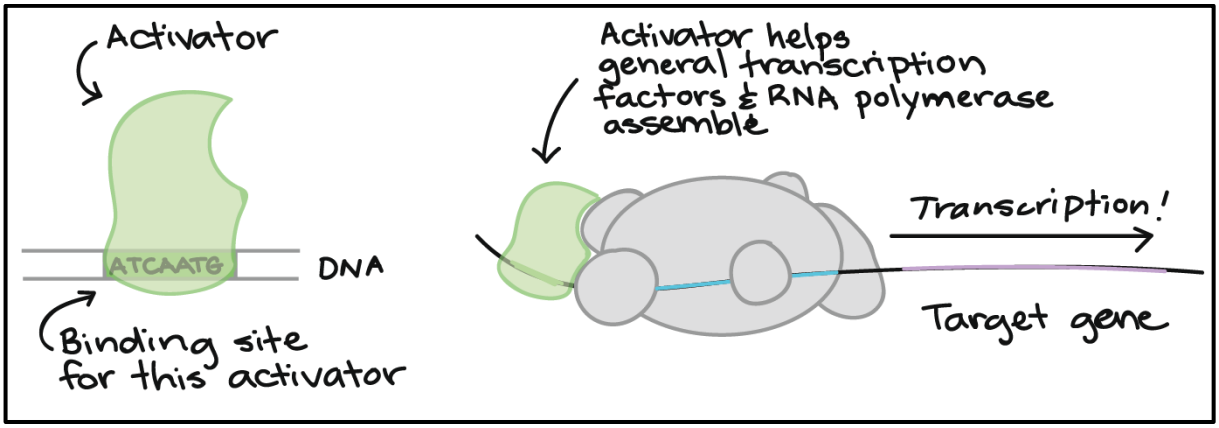


Fig: An activator is shown coupled to a specific DNA sequence that serves as its binding site.

The unbound end of the transcriptional activator interacts with general transcription factors, assisting polymerase and general transcription factors in assembling at the adjacent promoter.

Repressors:

Other transcription factors inhibit gene expression. This repression might take many different forms. A repressor, for example, may obstruct the binding of basal transcription factors or RNA polymerase to the promoter, preventing them from starting transcription.

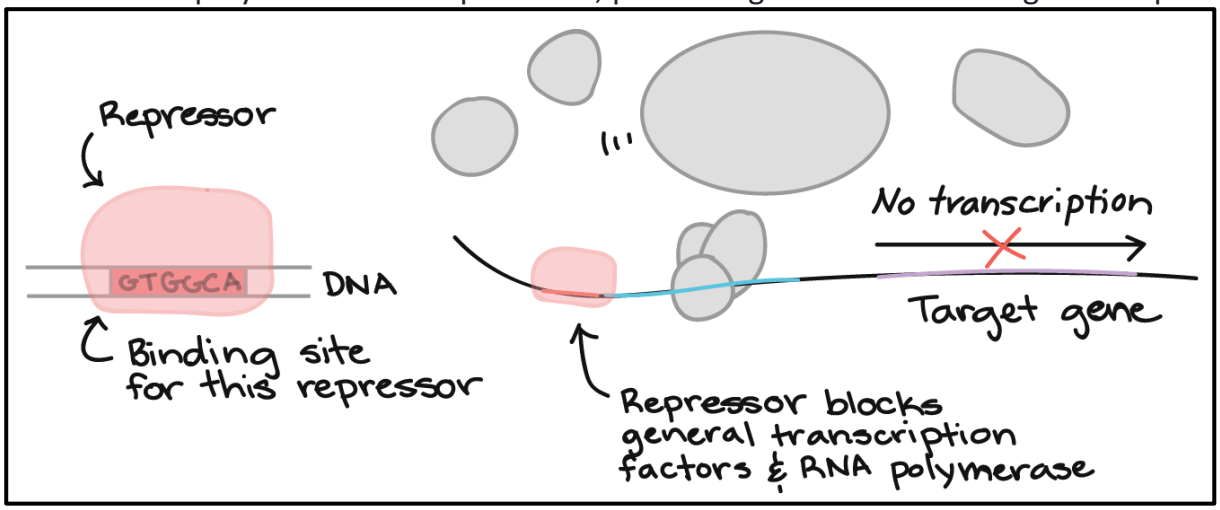
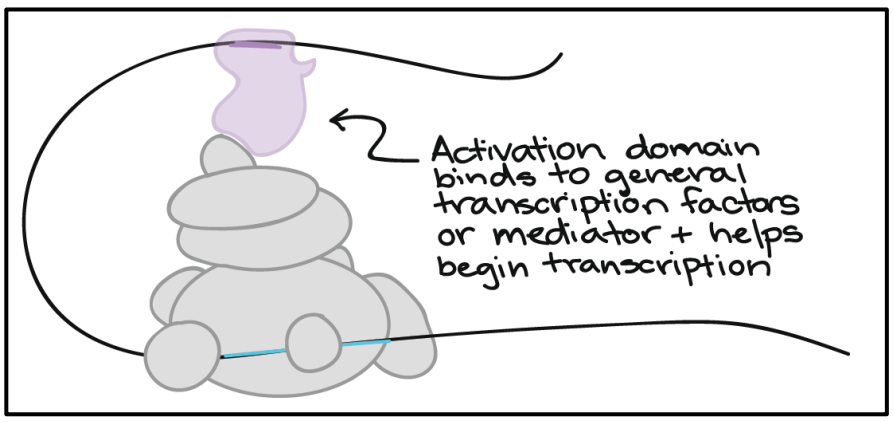


Fig: A repressor is shown connected to a specific DNA sequence that serves as its binding site in this diagram.

The repressor binds to this location and prevents the development of the transcription initiation complex at a nearby gene's promoter and still effect gene transcription

Binding sites: Transcription factor binding sites are frequently found around a gene's promoter. They can, however, be located in other sections of the DNA, sometimes distant



The DNA binding domain (which binds to the recognition site in DNA) and the activation domain (which is the "business end" of the activator that actually induces transcription, such as by promoting creation of the transcription initiation complex) are the two sections of an activator protein. The flexibility of DNA is what allows transcription factors to work at far-flung binding sites. The DNA loops act like cooked spaghetti, bringing distant binding sites and transcription factors closer to the general transcription factors or "mediator" proteins.

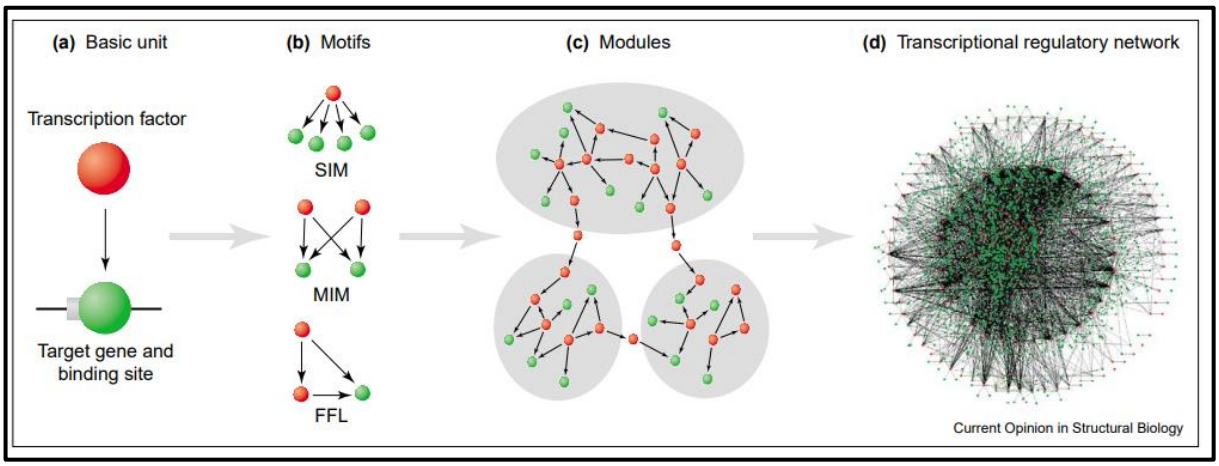
TRANSCRIPTIONAL REGULATORY NETWORK:

A directed graph can be used to represent the regulatory interactions between transcription factors and their target genes. These regulatory networks have a scale-free topology on a global scale, showing the presence of regulatory hubs. Substructures such as motifs and modules can be recognised at a local level in these networks. Despite the fact that networks spanning the evolutionary spectrum are organised similarly, there are interesting qualitative distinctions across network components, such as transcription factors. Despite the fact that the DNA-binding domains of transcription factors expressed by a given organism are drawn from a small number of historically conserved superfamilies, their relative abundance varies dramatically among phylogenetic groups. Many of these networks appear to have evolved through substantial duplication of transcription factors and targets, with regulatory interactions from the ancestor gene typically being passed down. Interactions are conserved in different ways across genomes. The structure and evolution of these networks can be translated into predictions that can be utilised to engineer different organisms' regulatory networks.

The transcriptional regulatory network's structure:

As depicted in the picture, the assembly of regulatory interactions linking transcription factors to their target genes in an organism can be regarded as a directed graph, with regulators and targets as nodes and regulatory interactions as edges. This network is a multi-layered, sophisticated structure that may be investigated at four different degrees of detail. The network, at its most basic level, is made up of transcription factors, downstream target genes,

and DNA binding sites, as shown below. At the second level, these fundamental units are organised into network motifs, which are recurring patterns of linkages that emerge often across the network. At the third level, the motifs form modules, which are transcriptional units that are semi-independent. Finally, the regulatory network at the top level is made up of interconnected interactions amongst modules that make up the complete network. It's worth noting that most regulatory network research has concentrated on Escherichia coli and the yeast Saccharomyces cerevisiae, which have the greatest data. Individual regulatory interactions in E. coli were painstakingly gathered from the literature and entered into the RegulonDB database. In yeast, however, the output of large-scale DNA-binding data from chromatin immunoprecipitation-chip (ChIp-chip) assays has considerably supplemented manually vetted data.



Structural organisation of transcriptional regulatory networks:

(a) The transcription factor, its target gene with DNA recognition site, and the regulatory interaction between them make up the 'basic unit.'

(b) Units are frequently organised into network ‘motifs,' which are over-represented patterns of inter-regulation in networks. Single input (SIM), multiple input (MIM), and feed-forward loop (FFL) motifs are examples of motifs.

(c) Network motifs can be combined to construct semi-independent ‘modules,' several of which were discovered by combining regulatory interaction data with gene expression data and applying evolutionary conservation.

(d) The 'transcriptional regulatory network,' which provides the pattern for gene expression regulation in an organism, is made up of all regulatory interactions.

GRAPH MINING FOR TRN:

Graphs are ideal for simulating complicated structures found in the actual world. Gene regulatory networks, for example, are directed graphs with genes as nodes and regulatory impacts as connections. A directed graph with nodes representing pages and edges representing hyperlinks is a good model for the World Wide Web. Directed graphs are also suitable for representing email conversations, social graphs in which an individual may follow the actions of others, and citation networks in which an article cites other papers.

Graphs are intensively researched in graph theory and, more recently, in the context of data mining, due to their numerous uses. The majority of studies focus on either unlabeled graphs or graphs with unique labels associated with each node. Labels in a social graph, for example, can be people's IDs; in a gene regulatory network, they can be genes' names; in a Web graph, nodes are frequently labelled with page URLs; and so on.

Objects (represented by nodes) can, however, be connected with a variety of attributes in many applications. Individuals in social graphs, like genes in regulatory networks, have a variety of features. Articles in citation networks are linked to a variety of data, including keywords, authors, publication dates, patents, and so on. Each attribute connected with an object is an attribute of the appropriate node in these data. Attributed graphs are graphs in which nodes are annotated with sets of attributes (or itemsets), and only a few research have been dedicated to their study.

The finding of frequent subgraphs is a common task while examining graph data. Such patterns are interesting because they demonstrate how node labels are frequently structured. Mining attributed graphs reveals structural trends while also highlighting the link between node characteristics.

The mining of attributed graphs is difficult for two reasons. To begin, you must combine network structure analysis with the discovery of frequent itemsets connected with nodes. The second reason is that, as with labelled graphs, the cost of subgraph isomorphism tests has a significant impact on the mining process' performance.

However, this issue has received little attention to far. Indeed, in labelled graphs, nodes are associated with distinct labels, which can be numerous. As a result, having the same label linked with many nodes in a subgraph is uncommon. In attributed graphs, this is different because entities represented by nodes have several properties, some of which are quite common. As a result, the number of possible automorphisms is considerably increased. In a social graph, for example, a given individual is often connected with a group of other individuals sharing several characteristics (age group, hobbies, social class, musical taste, etc.).

Constructing transcriptional regulatory networks:

Approaches are devised to evaluate (1) the identity and expression level of interacting nodes, (2) how interactions change with time (e.g., through a cell cycle or during differentiation), and (3) the phenotypic impact of disrupting key nodes in order to understand the topology and dynamics of transcriptional regulatory networks governing biological processes such as the cell cycle or differentiation. The eukaryotic transcriptional regulation machinery's complexity reflects the wide range of responses it regulates, making elucidating transcriptional regulatory networks a tough undertaking. This raises obvious issues about how a specific transcriptional response is elicited, such as how a signalling pathway activates a specific TF, how temporal specificity is formed, and where target specificity comes from. As a result, accurately accounting for all layers of regulation is now difficult, if not impossible, and some assumptions are made. For example, it is commonly considered that the steady-state level of an mRNA (as determined by DNA microarrays in an expression profiling experiment) represents the rate of transcription or the amount of protein translated from that mRNA. Furthermore, it is frequently assumed that if a transcription factor is expressed, it is active,

despite the fact that dimerization, post-translational changes, subcellular localization, and other aspects must all be taken into account.

CHAPTER 3

METHODOLOGY

3.1 Transcription Adjustment Network

The distribution and structure of the network, the distribution of degrees and its logarithmic scale are shown in Figures 5 and 6. According to Figures 5 and 6, the shape of the network structure toScale Free is more similar Especially in Figure 6. Also in Figure 5, if the first and last nodes are ignored, a linear function is observed that indicates the distribution of the network.

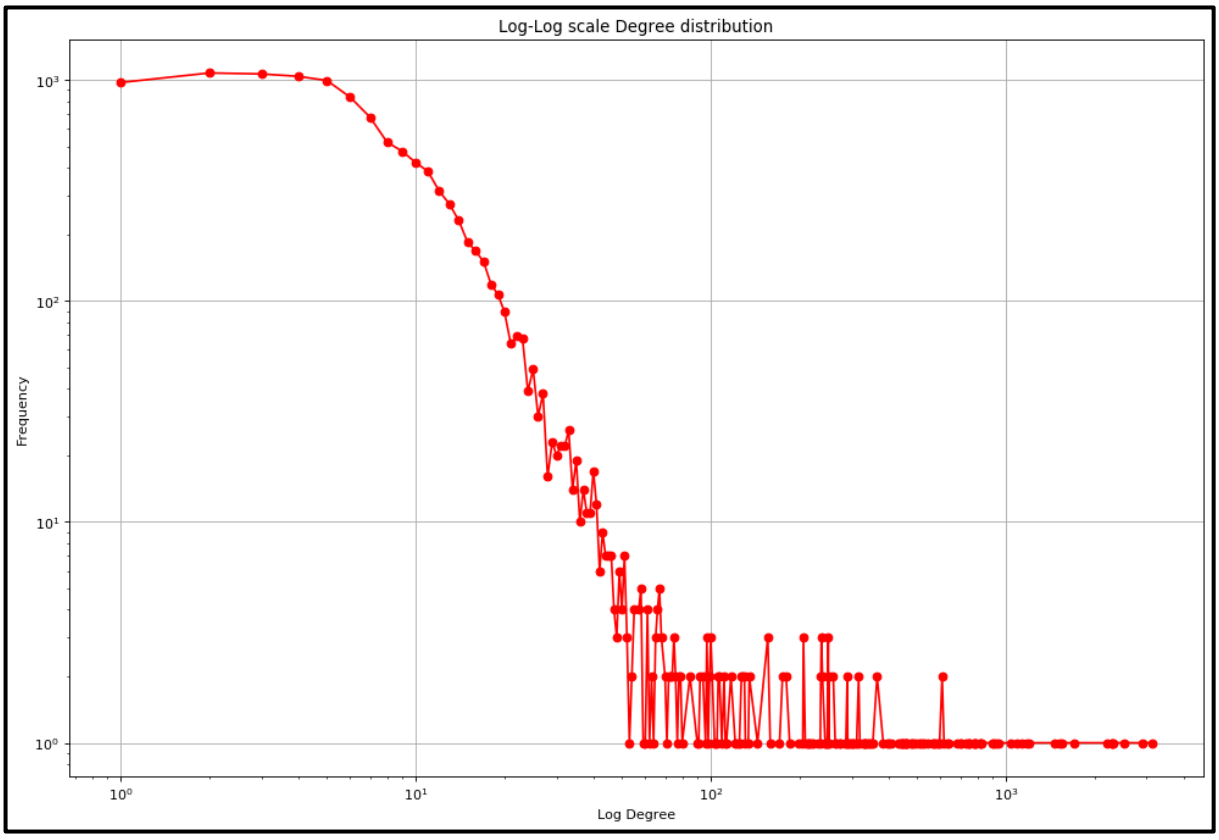


Figure 5: Distribution of network degrees on a logarithm-logarithm scale

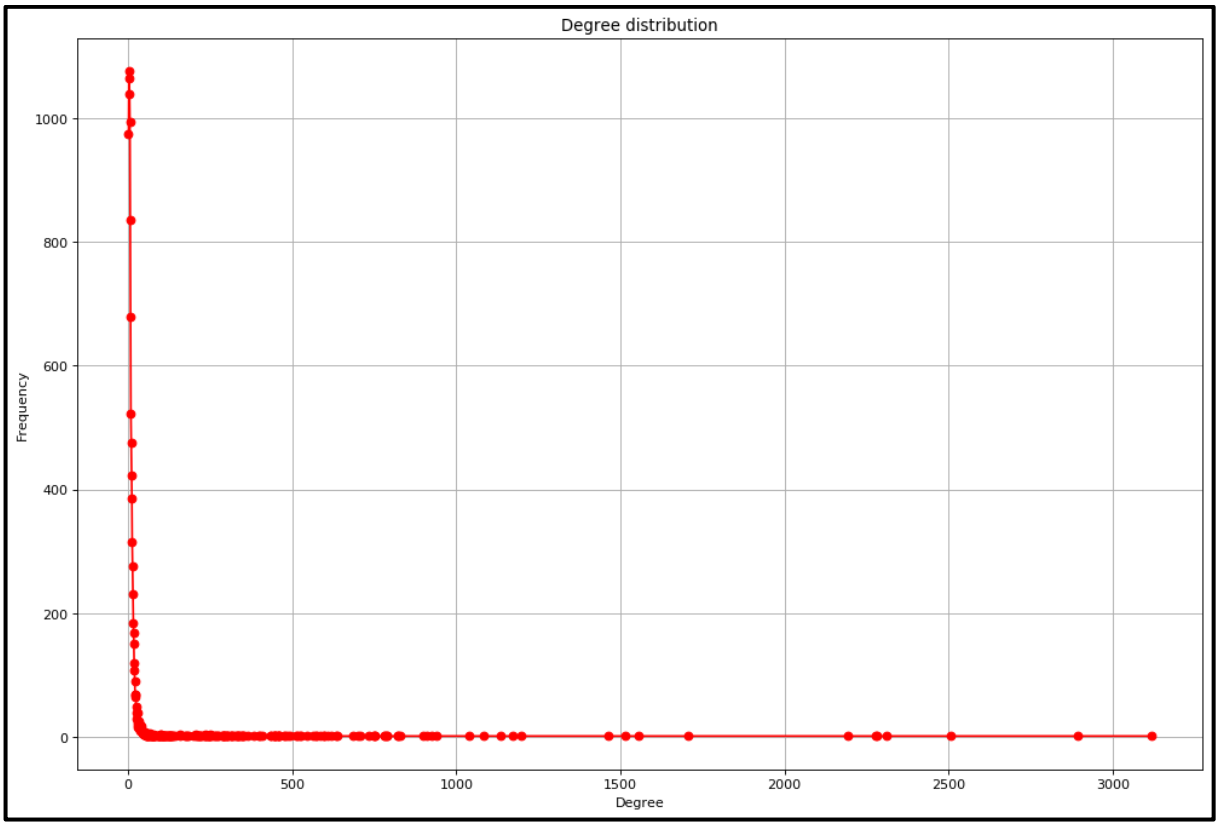


Figure 6: Distribution of network degree

For random network	For the desired network	
3,367	0.297	Average distance
0.0014	0.223	Average clustering coefficient

Table 7: Shows that the existing network is far from random

Given that the distribution network, by adding a new node to the network, a new node is expected to have the node with the highest Scale Free Communication and degree is absorbed and connected. Also, nodes with lower rank in the network have more frequency, which seems to be more important in the network. Therefore, they are considered suspicious options for being the cause of lung cancer, which are further used by other criteria, more and are examined more closely.

3.2 Check network centre's

Three-degree centers, proximity and medianity are calculated for this network and the common nodes between at least two centers and more are important. There are many in the network because if cancerous mutations occur in them, because they are in the path of more genes or are associated

with, they are more with other nodes, they also have a shorter path with other nodes and it is less expensive to reach these nodes. Many genes are affected. Therefore, these nodes (with more centrality) are considered as genes that cause lung cancer. At The continuation of the algorithm used is described in more detail.

Centrality of proximity	Centrality of Miandari	Degree degree	
0	0	0.00009	minimum amount
0.0295	1.86 *5-10	0.0014	Average value
0.54	0.009	0.283	Maximum amount
MYC	MYC	MAX	Node corresponding to the maximum value

Table 8: Information about network centers

Here, two algorithms are used to find the genes that cause cancer.

First algorithm:

1. Calculate the degree of centrality, proximity, and median for all nodes

2. Calculate the average for all three centers

3. Define the rounded value of the average as a threshold, i.e., for the centers named in line 1, 002.0, 0.000019 and 0.03, respectively. (The amount of rounding is proportional to the scale of each performed)

4. Find and separate data with more centralization than the threshold separately in each centralization

5. Find and separate common data available in all three centre's

6. Filter on real cancer data and delete data that is not in the current data set

7. Label the actual data obtained in line 6 with the number 1

8. Label the predicted data expressed in line 5 with the number 1 if present in line 6 and 0 if not present in line 6

9. Equalize the number of actual and predicted data using the number 0

10. Comparison of real and predicted data and evaluation of the model using the criteria of accuracy, call, accuracy and value -F *.

The second algorithm has the same steps as the first algorithm, except for line 5, which instead of considering the common data of all three centers, from data that exists in only two centers is used. According to Table 8 and the second algorithm, nodes. It is an important node and if it has a trend mutation, it is effective in the network. So, this is MYC The node is thought to be the gene that causes lung cancer.

CHAPTER 4

RESULT AND DISCUSSION:

According to line 5 of the first algorithm, the nodes that are present in all three parts i.e., betweenness centrality, degree centrality, closeness centrality and are common to all three are used as cancer-causing genes. The lungs chose 1015 nodes with the highest centrality, 492 nodes with the highest center of gravity, and 772 nodes with the highest degree of centrality were found and among them, there are 427 common nodes that are considered as cancer-causing genes and are shown in Table 10. Are given. Then compare and evaluate them with the actual cancer-causing genes in the data set (493 nodes).

	Name
2	ABL1
3	STAT5B
4	RB1
5	STAT5A
6	MYC
7	APBB1
8	APC
9	CTNNB1
10	AR
11	CITED2
12	ELK1
13	ESR1
14	ESR2
15	ETS1
16	FHL2
17	FOXA1
18	GATA3
19	GTF2F1
20	HES1

Table 10: Sample of 20 cancer-causing genes found with the highest number of centers with the first algorithm.

```
confussion matrix Alg1:
[[113   0]
[399  94]]

clasification report for Alg1:
              precision    recall  f1-score   support

           0       0.22      1.00      0.36       113
           1       1.00      0.19      0.32       493

    accuracy                           0.34       606
   macro avg       0.61      0.60      0.34       606
weighted avg       0.85      0.34      0.33       606
```

Table 11: Integration matrix and evaluation values of the first algorithm for each class

```
Accuracy1: 0.3415841584158416
Accuracy2: 0.3564356435643564
F1 score1: 0.32027257240204426
F1 score2: 0.375
Recall1: 0.19066937119675456
Recall2: 0.23732251521298176
Precision1: 1.0
Precision2: 0.8931297709923665
```

Table 12: Evaluation values for the first and second algorithm

Also, using the second algorithm, 606 common nodes are created, which are considered as cancer-causing genes, and Labelling was performed for 493 actual data with 1 and for 113 non-cancerous residual data with 0. Evaluation of these factor genes the predicted cancer can be seen in Tables 12 and 13.

```
 confussion matrix Alg2:
...
    accuracy                           0.36       606
   macro avg       0.55      0.56      0.36       606
weighted avg       0.77      0.36      0.37       606
```

Table 13: Integration matrix and evaluation values of the second algorithm for each class.

According to the first algorithm, 113 cancer-causing genes and 94 non-cancer-causing genes were correctly predicted, and 399 were cancer-causing genes, but with Due to the wrong algorithm, non-

cancerous genes are predicted. Also, no non-cancerous data using this algorithm, to the cause of the cancer is not predicted.

On the other hand, according to the second algorithm, 99 cancer-causing genes and 117 non-cancer-causing genes were correctly predicted, and 376 were cancer-causing genes. But according to the algorithm, the non-cancer gene and the 14 non-cancer data are incorrectly predicted. In addition, according to Tables 11 and 13, it can be concluded that the first algorithm in detecting non-active lung cancer genes and the second algorithm in Detection of lung cancer genes works better. In general, the degree of accuracy, recall and the value of the algorithms indicates that the second algorithm works better than the first algorithm and to build-F A program for classifying genes It is better to use the second algorithm.

Network socialization:

For socialization, a connected network is used and single observation as remote data that correspond to the non-functional gene ZNF471 It is cancer and it does not matter in this matter, and as mentioned earlier, the network is considered directional networks. Therefore, for socialization, the network must change from directional to directionless. Using the greedy algorithm, Levin and Publish The label deals with the socialization of the network without direction and the results are as shown in Table 14.

Label Diffusion Algorithm	Levin's algorithm	Greedy algorithm	
3	13	7	Number of communities found
05-e4,803	0.306	0.295	The modularity value

Table 14: Information from socialization

According to the modulus value shown in Table 14, the two greedy and Levin algorithms work almost well and the resulting communities have the nodes are dense and well away from random.

Using this algorithm, the influential genes (propagation occurs in them) are identified in each community, and finally the total of the influential genes. The results from all communities were predicted as lung cancer genes and evaluated using certain criteria. To choose a better algorithm, a comparison of speed and accuracy is used here (Levin and Greedy).

Given the type of data, if we want to find a group that includes cancer-causing genes, it is better to cluster rather than socialize. Because the goal is to find groups with similar characteristics, not dense groups (necessarily between the genes that cause cancer. Does not exist), and to perform clustering requires characteristics of genes that can be used to cluster. But the main purpose of socializing in this project, rather than socializing to include cancer-causing genes, is to find dense communities and it is suitable to be used in the maximum diffusion algorithm.

References

1. https://academic.oup.com/bioinformatics/article/23/13/i577/237816?login=false
2. https://ieeexplore.ieee.org/abstract/document/1544466
3. https://www.pnas.org/doi/10.1073/pnas.1702581114
4. https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0168760.g002
5. https://www.embopress.org/doi/full/10.15252/msb.20167435
6. https://www.frontiersin.org/articles/10.3389/fphys.2016.00568/full
7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4943455/
8. https://pubmed.ncbi.nlm.nih.gov/22895549/
9. https://link.springer.com/chapter/10.1007/978-3-319-96136-1_27
10. https://ieeexplore.ieee.org/document/1635809?denied=
11. https://www.ijbiotech.com/article_143741_08403f57cda156ef675ed9b31e79542f.pdf